

METHOD AND APPARATUS FOR DETERMINING MOLECULAR CRYSTAL STRUCTURES

The present invention provides an improved method and apparatus for determining molecular crystal structures from powder diffraction data. In particular, the present invention enables molecular crystal structures to be identified using only powder diffraction data in a manner that is considerably faster than is currently the case. Furthermore, with the present invention the molecular crystal structure of large organic molecules, such as pharmaceutical compounds, can be determined using data from powder diffraction analysis.

Information on the molecular crystal structure of a molecule is usually obtained through irradiation of a single crystal of the molecule with neutrons or X-rays. Subsequent analysis of the resultant diffraction pattern, which consists of a series of angularly spaced intensity peaks with each peak representing an individual Bragg reflection, provides information on the structure. Whilst this single crystal diffraction technique is an effective technique for determining the crystal structure of a molecule, it can often prove difficult to grow the single crystals necessary for the analysis. Moreover, where the molecule can crystallise in more than one polymorphic form, it is sometimes the case that it can prove very difficult to grow a single crystal of a particular polymorph.

To address these problems, a powder diffraction analysis was developed in which a crystalline powder of the material under analysis is irradiated instead of a single crystal. Analysis of the resultant diffraction pattern is hampered by the fact that the diffraction pattern may include Bragg reflections that partially or fully overlap one another, making it difficult for individual reflections to be identified, and their associated intensities quantified. An example of experimental data from irradiation of a powder sample of the drug substance cimetidine in the form of a graph representing intensity of the Bragg reflections with respect to angular position is shown in Figure 1. Currently, in order to identify molecular

The present invention provides a method for determining molecular crystal structures from powder diffraction data comprising the steps of: generating a reduced representation of the powder diffraction pattern in dependence on a predetermined unit cell and space group of the molecule under examination in which the total quantity of diffraction data is significantly reduced whilst maintaining the characteristics of the diffraction data that are representative of the crystal structure under examination; determining a set of variables for describing trial molecular structures, derived from predetermined internal co-ordinates and said space group; assigning values to said variables thereby creating a population of trial structures each defined by a unique set of values for said variables; calculating a fitness for each trial structure with respect to the reduced representation of the powder diffraction pattern; determining whether any

one of the calculated fitnesses is less than or equal to a predetermined threshold; where none of the calculated fitnesses is less than or equal to the threshold value, selecting at least one survivor from the population of trial structures, altering the values of the variables of at least one of the

5 survivors in accordance with one or more predetermined rules, calculating the fitnesses of the new trial structures; and repeating the steps of selecting survivors, altering the values of the variables and calculating the fitnesses of the new trial structures until at least one of the calculated

10 fitnesses is less than or equal to the threshold value, and where at least one of the calculated fitnesses is less than or equal to the threshold, outputting at least one trial molecular crystal structure represented by the successful sets of values.

In a further aspect the present invention provides apparatus for determining molecular crystal structures comprising a structure factor

15 analyser for generating from experimental powder diffraction data for the molecule under examination a reduced representation of the powder diffraction pattern based on a predetermined unit cell and space group in which the total quantity of diffraction data is significantly reduced whilst the characteristics of the diffraction data representative of the crystal structure

20 under examination are maintained; a controller for determining a set of variables for describing trial molecular structures, derived from predetermined internal co-ordinates and said space group; a searching processor for creating a population of trial structures each defined by a unique set of values for said variables said searching processor including a

25 fitness analyser for calculating a fitness for each trial structure with respect to the reduced representation of the powder diffraction pattern, a thresholding device for determining whether any one of the calculated fitnesses is less than or equal to a predetermined threshold, a survivor selector for selecting at least one survivor from the population of trial

30 structures, a variable adjustment device for altering the values of the variables of at least one of the survivors and output means for outputting the one or more trial molecular crystal structures having calculated

fitnesses less than or equal to the threshold value.

The reduced representation preferably consists of identification of each reflection in the powder diffraction data along with associated weighting factors and ideally is in the form of a structure factor intensity listing and associated covariance matrix. Reference herein to reflections is intended as reference to individual Bragg reflections or peaks in the diffraction data resulting from reflections of the incident radiation from the structure of the molecule.

Moreover, preferably, the fitness χ^2 of each of the trial structures is determined using the following function:

$$\chi^2 = \sum_h \sum_k \{ (I_h - c|F_h|^2) (V^{-1})_{hk} (I_k - c|F_k|^2) \}$$

where:

$I_{h,k}$ = extracted intensity from the structure factor analyser

V_{hk} = covariance matrix from the structure factor analyser

c = a scale factor

$F_{h,k}$ = calculated structure factor from trial structure

The plurality of co-ordinates preferably consist of three co-ordinates representative of the location of the molecule within the unit cell, three co-ordinates representative of the orientation of the molecule within the unit cell and one or more co-ordinates representative of respective torsion angles.

In a preferred embodiment the structure factor analyser additionally automatically determines the optimal unit cell and space group for the molecule under examination instead of the unit cell and space group being predetermined manually and input into the structure factor analyser. More preferably a co-ordinate generator is provided for automatically determining the set of internal co-ordinates instead of the set of internal co-ordinates being predetermined manually and input into the controller.

The search for a three dimensional structure of a molecule which

would produce a powder diffraction pattern nearly identical to available experimental results is greatly simplified with the present invention and considerably speeded up by a factor of typically up to 1000 or more. This is achieved by the reduction of trial molecular crystal structures to a unique
5 set of variables that does not explicitly include the co-ordinates of individual atoms in the molecule but instead includes a single set of co-ordinates for the location and orientation of the entire molecule. Reduction of the experimental powder diffraction data to data representative of each reflection in combination with a weighting factor also assists in speeding up
10 the analysis of trial crystal structures in which the fitness of each trial structure with respect to the reduced experimental data is determined using structure factor analysis.

The present invention relies on the fact that at its most basic, a molecular crystal structure can be represented by a set of internal co-
15 ordinates describing the molecule under investigation together with co-ordinates describing the location and orientation of the molecule within a unit cell of which only some but not all need be variable. The reduction of the molecular crystal structure to such a set of variables enables analysis of the trial structures to be performed much more quickly than an analysis
20 performed using the conventional method of describing the crystal structure in terms of the fractional or Cartesian co-ordinates of every atom in the asymmetric unit of the structure. Such conventional representations are considered to be unworkable in a model building sense because of the computing power necessary to position individual atoms independently of
25 each other.

The representation of the trial structures used in the invention along with the novel fitness function means that analyses can be performed in seconds or minutes on the current generation of conventional personal computers or workstations. Moreover, the representation is versatile as it
30 allows the invention to work with flexible molecules as well as multiple fragments.

An embodiment of the present invention will now be described by

Figure 1 is a graph of experimental data from x-ray powder diffraction analysis of cimetidine showing 315 reflections, using an irradiation wavelength of 1.5285Å and a data range for 2θ of 8°-56°;

Figure 2 is a schematic representation of the 2D molecular structure of cimetidine;

Figure 4 is a diagram of the crystal structure of cimetidine;
10 Figures 5a, 5b, 5c and 5d are diagrams showing the progressive development of a trial crystal structure for cimetidine, employing the method and apparatus in accordance with the present invention, overlying the diagram of Figure 4;

Figures 7a, 7b and 7c show the molecular structure of dopamine deuterobromide, a graph of the development of a trial structure for the crystal and a diagram of the solution respectively, employing the method and apparatus of the present invention;

Figure 9 is a tabulation of part of the reduced representation of the
 25 powder diffraction data of Figure 8 generated in accordance with the
 method of the present invention; and

The present invention will be described with reference to an experimental determination of the crystal structure of the molecule

cimetidine, a histamine H₂ antagonist used in the treatment of stomach ulcers, for which a full single crystal structure (monoclinic Form A) determination has already been performed. Figure 2 shows the 2D chemical formula of the cimetidine molecule, whilst the known arrangement of the cimetidine molecules within the unit cell of the crystal structure is shown in Figure 4.

To determine the molecular crystal structure of cimetidine employing the method and apparatus of the present invention with reference to Figure 3, initially a conventional powder diffraction pattern (10) is obtained from a crystalline powder sample of cimetidine. The resultant diffraction pattern is shown in Figure 1. The experimental diffraction data (10) is input into a cell dimension analyser (12). The cell dimension analyser (12) uses conventional techniques to assess the diffraction pattern in order to determine the unit cell dimensions of the crystal structure. The generation of the unit cell dimensions may alternatively be performed manually, however, it is preferred that the unit cell dimensions be generated automatically using the crystal modelling apparatus. The diffraction pattern is also input to a structure factor analyser (14) that also receives the unit cell dimensions determined by the analyser (12). The structure factor analyser (14) analyses the experimental diffraction pattern using the lowest symmetry space group consistent with the crystal class determined by the cell dimension analyser (12), reducing the data to a first structure factor intensity listing and an associated covariance matrix. From this listing, the true space group (16) of the crystal structure is determined and used by the structure factor analyser (14) to generate a second structure factor intensity listing and associated covariance matrix (18). By generating this second structure factor intensity listing and associated covariance matrix (18), the total quantity of the original experimental diffraction data is significantly reduced in amount without loss of those characteristics of the data representative of the crystal structure under examination. The original data can be reduced by a factor typically equal to the number of data points in the original powder diffraction data divided by the number Bragg

5

10

25

30

Preferably, the only unknown factors and so the only variables to be found in the internal co-ordinates are the values of the variable torsion angles (represented by variables $\tau_1, \tau_2, \tau_3, \tau_4, \tau_5 \dots$). It is not essential for the bond lengths and bond angles to be held fixed and where appropriate these factors too may be varied in determining the crystal structure of the molecule. It has been found though that variation of the bond lengths and bond angles within chemically sensible bounds has a much smaller effect on the calculated diffraction data than variation of the flexible torsion angles within the structure. Thus for most purposes, acceptable results can be achieved with these factors held fixed.

The output (24) of the co-ordinate generator (22) is supplied to a controller (26) that is also connected to the space group output (16) of the structure factor analyser (14). The controller (26) also includes an input (28) to enable manual setting of selected operational parameters such as the number of trial structures to be analysed in each generation, i.e. the population size. The controller (26) uses the internal co-ordinates and the space group to determine additional variables representing the location and orientation of a molecular structure in the unit cell. Preferably, the location of the molecular structure within the unit cell is defined using a single reference point in fractional co-ordinate space represented by external co-ordinates or variables (x, y, z). The orientation of the molecule at that point may be described using Euler angles (α , β , γ). Alternatively, the orientation of the molecule may be described using a quaternion, q.

$$\{x, y, z, \alpha, \beta, \gamma, \tau_1, \tau_2, \tau_3, \tau_4, \tau_5 \dots\} \text{ or } \{x, y, z, q, \tau_1, \tau_2, \tau_3, \tau_4, \tau_5 \dots\}.$$

These variables are suitable for iterative mathematical processing and are more amenable to search procedures than the full complement of

individual atomic co-ordinates used in conventional techniques.

The output (30) from the controller (26) is then supplied to an iterative searching processor (32). The output (30) consists of the set of variables determined by the controller (26); the complete internal co-ordinates produced by the co-ordinate generator (22); operating parameters such as the selected size of the population to be employed in the searching procedure; any rules restricting or controlling the values which can be allocated to each of the variables; and any rules controlling the selection of survivors, the breeding and the mutation of survivors, described in greater detail below.

In the method shown in Figure 3, the iterative searching processor (32) employs a genetic algorithm to determine the correct molecular crystal structure. The above mentioned set of variables $\{x, y, z, \alpha, \beta, \gamma, \tau_1, \tau_2, \tau_3, \tau_4, \tau_5 \dots\}$ or $\{x, y, z, q, \tau_1, \tau_2, \tau_3, \tau_4, \tau_5 \dots\}$ are thus equated to chromosomes, with each individual variable equating to a gene. The genetic algorithm establishes certain protocols based on the concept of 'survival of the fittest', with respect to the selection of survivors, the breeding and the mutation of survivors.

Firstly, within the searching processor (32) an initial population of chromosomes is created (33) by assigning random numbers to each of the genes of each of the chromosomes. The allowable random numbers for any particular gene may be restricted in accordance with rules input from the controller (26). The selected size of this initial population depends somewhat upon the complexity of the structure under investigation, with larger population sizes typically being required for problems involving more variables. In the case of cimetidine, where seven torsion angles were allowed to vary, resulting in thirteen degrees of freedom, a population size of 150 was chosen. The fractional co-ordinates (x, y, z) and Euler angles are randomly set as real numbers normally bounded by the Euclidian normalisers of the space group. The variable torsion angles (τ) are typically randomly set as real numbers in the range 0° - 360° .

Using the internal co-ordinates a three dimensional structure of the

5

$$\chi^2 = \sum_h \sum_k \{ (I_h - c|F_h|^2) (V^{-1})_{hk} (I_k - c|F_k|^2) \}$$

10

 $I_{h,k}$ = extracted intensity from the structure factor analyser

V_{hk} = covariance matrix from the structure factor analyser

c = a scale factor

$F_{h,k}$ = calculated structure factor from trial structure

15

25

30

5

10

20

25

30

identified. If one of the offspring chromosomes has a fitness value which is less than or equal to the threshold value, or if a predetermined maximum number of generations has been reached, then the search procedure is stopped (43). On the other hand, if the fitness functions of the

5 chromosome offspring all exceed the threshold value and the counted number of generations is less than the maximum allowed number, then the offspring are returned for the selection of survivors (47) and for the creation of new offspring (51).

Additional rules may also be employed where appropriate to

10 constrain the allowable values for the variables. These rules are determined by the controller (26) that may utilise data on crystal fragments stored in a memory (53). For example, the controller (26) may search through stored crystallographic databases of known crystal structure fragments related to the molecule to provide prior information about torsion

15 angle values likely to be adopted by the structure. Such information can then be implemented either as hard limits on the allowable values the torsion angles may adopt, or as probability distributions for the torsion angles. Furthermore, fragments of the molecule may be located using Patterson methods or direct methods. For example, the location of a

20 heavy atom may be used to anchor a molecule during the analysis by the searching processor (32). This effectively reduces the dimensionality of the problem by three as the fractional reference co-ordinates are then known.

Operation of the processor (32) in the search for the correct 3D

25 molecular crystal structure is thus an iterative procedure with the average fitness for each generation gradually tending towards the global minimum in fitness function space. In Figure 5a, a trial cimetidine crystal structure, corresponding to a chromosome in the first generation initialised at random by the processor (32), is shown overlying the true crystal structure first

30 shown in Figure 4. Figure 5b then shows one of the early offspring determined by the processor having a fitness value of $\chi^2=980$, again overlying the true crystal structure of cimetidine. In Figure 5c, a later

offspring having a fitness value of $\chi^2=430$ is shown and the improvement in correspondence between the trial crystal structure and the actual crystal structure is immediately evident. At this point, the crystal structure could be refined using a conventional constrained Rietveld refinement. Hence,
5 the processor (32) may be arranged so that the threshold value for the fitness function is set at around 450. This would result in the search procedure being stopped once the trial structure shown in Figure 5c had been generated, thereby enabling alternative methods to be used to refine the fine details of the trial structure. The advantage of stopping the search
10 procedure at this point is that, usually, conventional methods will be able to refine the fine details of the structure more efficiently than the presently described method and apparatus.

Continuing with the present method, in Figure 5d an offspring having a fitness value of $\chi^2=110$ is shown at which point the detail of the trial
15 structure is easily refinable. Figure 6 is a graph of trial results for cimetidine using the method described above showing the fitness value of offspring with respect to the number of generations for both average fitness and the best fitness. As can be seen, a refinable structure is obtained within a few hundred generations, and an easily refinable structure is
20 obtained around 3000 generations. This latter structure corresponds to an elapsed time of approximate 40 minutes, with the processor running on a single 175MHz R10000 Silicon Graphics™ workstation.

As further examples for the speed of this method, easily refinable structures for pyrene were determined in around 33 seconds, around 15
25 seconds for chlorothiazide and 36 minutes for Ibuprofen, with all calculations being performed on a single 200MHz Pentium Pro™ personal computer.

The above method and apparatus may also be used with molecular structures consisting of more than one fragment. As shown in Figures 7a,
30 7b and 7c an easily refinable structure solution for dopamine deuterobromide using neutron powder diffraction data was achieved in around only 4000 generations. This structure involves not only a dopamine

Whilst in the examples given above the individual genes are real numbers, they could equally be represented by binary strings or integer approximations with appropriate scaling factors. Also, in the example given above the experimental diffraction data is reduced to a structure factor listing and associated covariance matrix, it will be apparent that alternative ways of reducing the total quantity of diffraction data may be employed which, although providing less faithful representations of the diffraction data, nevertheless preserve sufficient characteristics of the original diffraction data to enable a successful structure determination to be performed. For example the data relating to the correlation of reflections may be omitted from the reduced representation.

15 In the above example a genetic algorithm searching processor is employed to perform an iterative selection of candidate molecular crystal structures. Alternative iterative analysis processes such as simulated annealing, evolutionary strategies and neural network analysis may be used instead of the genetic algorithm. For example, using a simulated
20 annealing process, only a single member is normally utilised and the same variables that were treated as genes by the genetic algorithm are individually adjusted by a small perturbation of their current values. If the function value (χ^2 as defined previously) is better than before, then the new values of the variables are retained. If the function value is worse, then the
25 new values of the variables are not automatically rejected. Instead the new values may be retained if allowed by the temperature dependent Boltzmann selection protocol. In this way, 'uphill' (in terms of χ^2) adjustment of the variables is permissible, helping the algorithm to escape from local minima. The initial choice of the temperature is usually high to
30 allow large 'uphill' moves if necessary, but the temperature is usually lowered in some predetermined fashion during the iterative process. One such way is a temperature reduction that cools more slowly if the χ^2

An example employing the simulated annealing process to the determination of the crystal structure of capsaicin is shown in Figures 8 to 10. Experimental powder diffraction data for capsaicin is shown in Figure 8 with the first page of a tabulation of the reduced representation of the data produced using the above mentioned method shown in Figure 9. In Figure 10 the crystal structure obtained from powder diffraction data alone is overlaid upon the crystal structure obtained by the conventional single crystal diffraction route is shown. The experimental powder diffraction data consists of 9901 data points whilst the reduced representation of this data as exemplified in Figure 9 contains only 379 data points, in total.

1 0 0 10.047 0.1106 1 0 0 0 0 0 0

is an isolated reflection whereas at lines 18 and 19 two correlated
30 reflections are identified. At lines 41 and 42 two reflections are identified
that lie so close together that they are treated as a single variable intensity.
With the reduced representation the diffraction data has been effectively

The close agreement between the two structures shown in Figure 10 demonstrates one characteristic of the method described above - the ability of the simulated annealing method to 'fine tune' the structure, generally finding a solution very close to the global minimum in χ^2 space. This structure solution, which involved the optimisation of 16 degrees of freedom (10 internal, 6 external) took approximately 40 minutes to execute on a DEC Alphastation 500/500.

20 In a further adaption of the method, regardless of the global optimisation strategy used, both local and semi-global optimisation methods (e.g. Newton-Raphson, simplex) can be invoked when the χ^2 value reaches some predetermined value that is anticipated to be in the vicinity of the global minimum, thus providing accelerated convergence.

25 With the method and apparatus described above, molecular crystal structures may be solved from powder diffraction data alone. Definition of the molecular fragments in terms of internal co-ordinates means that for a single molecular fragment, problem complexity scales with the number of variable torsion angles rather than with the number of atoms in the
30 fragment. Thus, complex structures can be represented by quite short chromosomes and solved relatively easily. The simple description of

To assist in an understanding of the invention, the method has been described with reference to functional, i.e. analyser/processor units. It will of course be apparent that in practice the method is implemented as a program on a computer. Indeed, one of the advantages of this method is that the program can be implemented on a number of different computer architectures, including personal computers and a network of personal computers/workstations acting as a parallel computer.

5 To assist in an understanding of the invention, the method has been described with reference to functional, i.e. analyser/processor units. It will of course be apparent that in practice the method is implemented as a program on a computer. Indeed, one of the advantages of this method is that the program can be implemented on a number of different computer
10 architectures, including personal computers and a network of personal computers/workstations acting as a parallel computer.

10 architectures, including personal computers and a network of personal
computers/workstations acting as a parallel computer.

項目	1990年	1991年	1992年	1993年	1994年	1995年	1996年	1997年	1998年	1999年	2000年	2001年	2002年	2003年	2004年	2005年	2006年	2007年	2008年	2009年	2010年	2011年	2012年	2013年	2014年	2015年	2016年	2017年	2018年	2019年	2020年	2021年	2022年	2023年	2024年	2025年	2026年	2027年	2028年	2029年	2030年	2031年	2032年	2033年	2034年	2035年	2036年	2037年	2038年	2039年	2040年	2041年	2042年	2043年	2044年	2045年	2046年	2047年	2048年	2049年	2050年	2051年	2052年	2053年	2054年	2055年	2056年	2057年	2058年	2059年	2060年	2061年	2062年	2063年	2064年	2065年	2066年	2067年	2068年	2069年	2070年	2071年	2072年	2073年	2074年	2075年	2076年	2077年	2078年	2079年	2080年	2081年	2082年	2083年	2084年	2085年	2086年	2087年	2088年	2089年	2090年	2091年	2092年	2093年	2094年	2095年	2096年	2097年	2098年	2099年	2100年																																																								
人口	120,000,000	121,000,000	122,000,000	123,000,000	124,000,000	125,000,000	126,000,000	127,000,000	128,000,000	129,000,000	130,000,000	131,000,000	132,000,000	133,000,000	134,000,000	135,000,000	136,000,000	137,000,000	138,000,000	139,000,000	140,000,000	141,000,000	142,000,000	143,000,000	144,000,000	145,000,000	146,000,000	147,000,000	148,000,000	149,000,000	150,000,000	151,000,000	152,000,000	153,000,000	154,000,000	155,000,000	156,000,000	157,000,000	158,000,000	159,000,000	160,000,000	161,000,000	162,000,000	163,000,000	164,000,000	165,000,000	166,000,000	167,000,000	168,000,000	169,000,000	170,000,000	171,000,000	172,000,000	173,000,000	174,000,000	175,000,000	176,000,000	177,000,000	178,000,000	179,000,000	180,000,000	181,000,000	182,000,000	183,000,000	184,000,000	185,000,000	186,000,000	187,000,000	188,000,000	189,000,000	190,000,000	191,000,000	192,000,000	193,000,000	194,000,000	195,000,000	196,000,000	197,000,000	198,000,000	199,000,000	200,000,000	201,000,000	202,000,000	203,000,000	204,000,000	205,000,000	206,000,000	207,000,000	208,000,000	209,000,000	210,000,000	211,000,000	212,000,000	213,000,000	214,000,000	215,000,000	216,000,000	217,000,000	218,000,000	219,000,000	220,000,000	221,000,000	222,000,000	223,000,000	224,000,000	225,000,000	226,000,000	227,000,000	228,000,000	229,000,000	230,000,000	231,000,000	232,000,000	233,000,000	234,000,000	235,000,000	236,000,000	237,000,000	238,000,000	239,000,000	240,000,000	241,000,000	242,000,000	243,000,000	244,000,000	245,000,000	246,000,000	247,000,000	248,000,000	249,000,000	250,000,000	251,000,000	252,000,000	253,000,000	254,000,000	255,000,000	256,000,000	257,000,000	258,000,000	259,000,000	260,000,000	261,000,000	262,000,000	263,000,000	264,000,000	265,000,000	266,000,000	267,000,000	268,000,000	269,000,000	270,000,000	271,000,000	272,000,000	273,000,000	274,000,000	275,000,000	276,000,000	277,000,000	278,000,000	279,000,000	280,000,000	281,000,000	282,000,000	283,000,000	284,000,000	285,000,000	286,000,000